

1 **CAPICE: a computational method for Consequence-Agnostic Pathogenicity Interpretation**  
2 **of Clinical Exome variations**

3 Shuang Li<sup>1,2\*</sup>, K. Joeri van der Velde<sup>1,2\*</sup>, Dick de Ridder<sup>3</sup>, Aalt D.J. van Dijk<sup>3,4</sup>, Dimitrios Soudis<sup>5</sup>, Leslie R.  
4 Zwerwer<sup>5</sup>, Patrick Deelen<sup>1,2</sup>, Dennis Hendriksen<sup>2</sup>, Bart Charbon<sup>2</sup>, Marielle van Gijn<sup>1</sup>, Kristin M. Abbott<sup>1</sup>, B.  
5 Sikkema-Raddatz<sup>1</sup>, Cleo C. van Diemen<sup>1</sup>, Wilhelmina S. Kerstjens-Frederikse<sup>1</sup>, Richard J. Sinke<sup>1</sup>, Morris  
6 A. Swertz<sup>1,2</sup>

7 <sup>1</sup>University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, the  
8 Netherlands

9 <sup>2</sup>University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen,  
10 the Netherlands

11 <sup>3</sup>Bioinformatics Group, Wageningen University & Research, Wageningen, the Netherlands

12 <sup>4</sup>Biometris, Wageningen University & Research, Wageningen, the Netherlands.

13 <sup>5</sup>Donald Smits Center for Information and Technology, University of Groningen, Groningen, the Netherlands

14

15 \*equal contribution.

16

17 Corresponding author:

18 Morris Swertz

19 E-mail: [m.a.swertz@rug.nl](mailto:m.a.swertz@rug.nl)

20

21 **ABSTRACT (<=100 words)**

22 Exome sequencing is now mainstream in clinical practice, however, identification of pathogenic  
23 Mendelian variants remains time consuming, partly because limited accuracy of current  
24 computational prediction methods leaves much manual classification. Here we introduce CAPICE,  
25 a new machine-learning based method for prioritizing pathogenic variants, including SNVs and  
26 short InDels, that outperforms best general (CADD, GAVIN) and consequence-type-specific  
27 (REVEL, ClinPred) computational prediction methods, for both rare and ultra-rare variants.  
28 CAPICE is easily integrated into diagnostic pipelines and is available as free and open source  
29 command-line software, file of pre-computed scores, and as a web application with web service  
30 API.

31

32 **KEYWORDS**

33 Variant Pathogenicity Prediction, Machine Learning, Exome Sequencing, Molecular  
34 Consequence, Allele Frequency, Clinical Genetics, Genome Diagnostics

## 35 BACKGROUND

36 The past decades have seen rapid advances in genetic testing and increasing numbers of trial  
37 studies aimed at using genetic testing to facilitate rare disease diagnostics, and many studies  
38 have demonstrated the unique role whole exome and genome sequencing can play in improving  
39 diagnostic yield [(1), (2), (3), (4), (5), (6), (7)]. However, the vast amount of genomic data that is  
40 now available has created large interpretation challenges that can be alleviated using  
41 computational tools. However, variant interpretation in particular still remains time-consuming, in  
42 part because of the limited accuracy of current computational prediction methods and the manual  
43 work required to identify large numbers of false positives produced by those methods [(8), (9),  
44 (10)].

45 Existing prediction methods can be categorized into two groups. One group of methods  
46 [(11), (12)] focuses on specific types of variants, with the majority of these methods only  
47 classifying non-synonymous single nucleotide variants (nsSNVs) [(13), (14)]. Successful methods  
48 of this group include Clinpred (15), which has the best current performance validated in multiple  
49 datasets, and REVEL [(16)], which specifically targets rare variants. However, these methods  
50 miss the diagnosis when the causal variant is not an nsSNV, which is the case for 76% of reported  
51 pathogenic variants (17). The other category of prediction methods provides predictions of  
52 selective constraints for a broader range of variations [(18), (19), (20), (21)] that can also inform  
53 pathogenicity classification. A method that is widely used and acknowledged for performance is  
54 CADD [(22)], which estimates the deleteriousness of SNVs and short insertions and deletions  
55 (InDels). However, these tools are built for estimating evolutionary constraints and do not directly  
56 target pathogenicity. They can also introduce ascertainment bias for variants that are under high  
57 evolutionary pressure (such as nonsense and splicing variants) even though these can be  
58 observed in healthy populations, and they can neglect rare and recent variants that have not  
59 undergone purifying selection but are still found to contribute to diseases [(23)].

60           New computational prediction methods need to be examined for their ability to reduce the  
61 number of variants that requires time-consuming expert evaluation as this is currently a bottleneck  
62 in the diagnostic pipeline. With hundreds to thousands of non-pathogenic variants identified in a  
63 typical patient with a rare genetic disorder, it is important to restrict the false positive rate of  
64 computational prediction methods, i.e. the number of neutral variants falsely reported as  
65 pathogenic. However, new methods are currently often not evaluated for their ability to recognize  
66 neutral variants. Indeed, a recent review (24) found that commonly used variant interpretation  
67 tools may incorrectly predict a third of the common variations found in the Exome Aggregation  
68 Consortium (ExAC) to be harmful. We speculate that this may be explained by the bias in training  
69 data selection because the neutral set used in different tools can be biased towards common  
70 neutral variants [(15), (25), (26)], which in practice means that the pathogenicity of rare and ultra-  
71 rare variants cannot be accurately estimated. Therefore, it is important to avoid bias in data  
72 selection and evaluate false positive rate of the prediction methods in clinical setting where rare  
73 and ultra-rare neutral variants are frequently encountered using neutral benchmark datasets [(27),  
74 (28)] and clinical data.

75           The challenge for rare disease research and diagnostics is thus to find robust classification  
76 algorithms that perform well for all the different types of variants and allele frequencies. To meet  
77 this challenge, we developed CAPICE, a new method for Consequence-Agnostic prediction of  
78 Pathogenicity Interpretation of Clinical Exome variations. CAPICE overcomes limitations common  
79 in current predictors by training a sophisticated machine learning model that targets  
80 (non-)pathogenicity, using a specifically prepared, high confidence and pathogenicity versus  
81 benign balanced training dataset, and using many existing genomic annotations across the entire  
82 genome (the same features that were used to produce CADD). In high quality benchmark sets  
83 CAPICE thus outperforms existing methods in distinguishing pathogenic variants from neutral  
84 variants, irrespective of their different molecular consequences and allele frequency and, to our

85 knowledge, CAPICE is the first and only variant prioritization method that targets pathogenicity  
86 prediction of all-types of SNVs and InDels, irrespective of consequence type.

87 Below we will describe the results of our performance evaluations, discuss features and  
88 limitations of our methodology, provide extensive details on the materials and methods used,  
89 concluding that CAPICE thus offers high accuracy pathogenicity classification across all  
90 consequence types and allele frequencies, outperforming all next-best variant classification  
91 methods. To make CAPICE easy to access, we have developed CAPICE as both a command-  
92 line tool and a web-app, and released it with pre-computed scores available as ready-to-use  
93 annotation files.

94

## 95 **RESULTS**

96 Below we report performance analysis of CAPICE compared to the best current prediction  
97 models using gold standard benchmark sets, analysis of the classification consistency of  
98 CAPICE across different allele frequency ranges and across different types of variants and a  
99 small practical evaluation where we applied CAPICE to a set of patient exomes.

100

### 101 **CAPICE outperforms the best current prediction methods**

102 CAPICE is a general prediction method that provides pathogenicity estimations for SNVs and  
103 InDels across different molecular consequences (Figure 1). In our performance comparison, we  
104 included recently published prediction methods and those that show best performance in  
105 benchmark studies. In case a tool was not able to provide a prediction we marked it as 'No  
106 prediction returned'. Because most prediction methods are built specifically for non-synonymous  
107 variants, we performed the comparison for both the full dataset and the non-synonymous subset.  
108 In our benchmark datasets, CAPICE performs as well or better than other current prediction  
109 methods across all categories (Figure 1, Supplementary Figure 3, Supplementary Figure 4,

110 *Supplementary Table 1, Supplementary Table 2*). We also examined the robustness of CAPICE's  
111 performance for rare and ultra-rare variants and variants that lead to different consequences.

112 For the full data, CAPICE outperformed CADD, the mostly used 'general' prediction  
113 method, and achieved an area under the receiver operating characteristic curve (AUC) of 0.89 as  
114 compared to 0.53 for CADD (shown in Supplementary Figure 3). For the non-synonymous subset,  
115 CAPICE outperformed all the other prediction methods and achieved an AUC of 0.97 (shown in  
116 Figure 1*b*). The majority of other methods we examined are built specifically for non-synonymous  
117 variants, with the exception of FATHMM-XF, which was developed for point mutations. For the  
118 non-synonymous subset, REVEL, which was built for rare variants, produced the second best  
119 result and achieved an AUC of 0.90.

120 To assess impact of these difference in practice we assumed a clinical setting with the aim  
121 to recognize 95% of the pathogenic variants (which is a very high standard in current practice).  
122 When using a threshold of 0.02 on CAPICE classification score, CAPICE correctly recognized 95%  
123 of pathogenic variants in the full test dataset and wrongly classified 50% of the neutral variants  
124 as pathogenic – which was the lowest number of misclassified variants among all the predictors  
125 we tested. In contrast, CADD with a score threshold of 20 achieved a comparable recall of 94%,  
126 but wrongly classified 85% of neutral variants as pathogenic. When using gene-specific CADD  
127 score thresholds based on the GAVIN method (29), the performance of CADD was better but still  
128 much worse than CAPICE. All other tested methods could give predictions less than 30% of the  
129 full dataset.

130 We also examined how well the prediction methods can recognize neutral variants in two  
131 neutral benchmark datasets. For both datasets, CAPICE's performance was comparable to or  
132 better than the current best prediction methods (Supplementary Table 2, Supplementary Table  
133 3).

Figure 1: CAPICE outperforms other predictors in discriminating pathogenic variants and neutral variants. a) True/false classification for all predictors tested against the full benchmark set that contains all types of variants. Top bar shows the breakdown of the test set. Other bars show the classification performance for each method. Purple blocks represent correct classification of pathogenic variants. Dark-blue blocks represent neutral variants. Pink and light-blue blocks denote false classifications. Gray blocks represent variants that were not classified by the predictor tested. Threshold-selection methods are in Method section. b) Receiver operating characteristic (ROC) curves of CAPICE with AUC values for a subset of the benchmark data that only contains non-synonymous variants (ROC curve for the full dataset can be found in Supplementary Figure 3). Each ROC curve is for a subset of variants displaying a specific molecular consequence. AUC values for the different methods are listed in the figure legend.

134

### 135 **CAPICE outperforms other current predictors for rare and ultra-rare variants**

136 CAPICE performs consistently across different allele frequencies and especially well for rare and  
137 ultra-rare variants. Here we repeated the evaluation strategy for the same benchmark dataset  
138 grouped into five allele frequency bins (Figure 2).

139 For the full benchmark dataset, CAPICE performed consistently above 0.85 of AUC for  
140 variants with an allele frequency <1%, while the performance of CADD version 1.4 (30), the  
141 current best method for indicating the pathogenicity of variants throughout the genome compared  
142 to LINSIGHT (31), EIGEN (32), DeepSEA (33) drops significantly in case of rare variants (Figure  
143 2a). For the non-synonymous subset, CAPICE consistently performed better than or comparably  
144 to the next-best method, REVEL, for variants within different allele frequency ranges, and better  
145 than all other methods (Figure 2b).

146 For common variants (defined here as having an allele frequency >1%), the number of  
147 available pathogenic variants was too small (14 pathogenic variants) to get an accurate and  
148 robust performance measurement.

Figure 2 Performance comparison for rare and ultra-rare variants (a) for variants of different  
molecular consequences (b) in the missense subset. Each dot represents the mean AUC value  
with standard deviation.

#### 149 **CAPICE shows consistent prediction performance for different types of variants**

150 CAPICE outperforms the current best computational prediction methods for variants that cause  
151 different molecular consequences (Figure 3 and Supplementary Figure 2). For variants displaying  
152 different molecular consequences, CAPICE has an AUC of 0.92 for canonical splicing variants  
153 and an AUC of 0.97 for non-synonymous variants in the independent test dataset. Compared to  
154 CADD, CAPICE performs significantly better for multiple types of variants, particularly canonical  
155 splicing, stop-gained and frame-shift variants.

Figure 3 Performance comparison for variants of different molecular consequences of CAPICE  
and CADD.

#### 156 **CAPICE performance in clinical setting**

157 In addition to the synthetic benchmark datasets, we also evaluated CAPICE's performance in  
158 patients' data.

159 To have first assessment of clinical utility, we used whole exome sequencing data from  
160 54 solved patients from our diagnostics department and compared the ranking of the disease-  
161 causing variant with scores from CADD and CAPICE. We did not compare to REVEL, the second  
162 best method from our previous evaluation because a specific method for non-synonymous  
163 variants can miss variants of other molecular effects. A description of the solved patients' can be  
164 found in (34). For each disease-causing variant discovered in that patient, we compared the



165 performance of CAPICE and CADD by comparing the ranking of the particular variant among all  
166 variants observed within that patient. For 83% of the cases, CAPICE can prioritize the disease-  
167 causing variant within the 1% of the total variants observed in whole exome sequencing  
168 experiment, while CADD achieves the 1% performance for only 60% of the cases. Consistent with  
169 results described in previous sections that CAPICE achieves better AUC value for frameshift  
170 variants, CAPICE performed better for all cases with a disease-causing variant of frameshift effect.

Figure 4 Performance comparison in real cases. In total, 54 patients and 58 variants were included.  
Each variant is reported as the diagnosis for that patient. Each dot in the plot shows a variant.  
The color of the dot represents the molecular effect predicted by VEP.

## 171 **DISCUSSION**

172 We have implemented a supervised machine learning approach called CAPICE to prioritize  
173 pathogenic SNVs and InDels for genomic diagnostics. CAPICE overcomes the limitations of  
174 existing methods, which either give predictions for a particular type of variants or showing  
175 moderate performance because they're built for general purposes. We showed in multiple  
176 benchmark datasets, either derived from public databases or real patient cases that CAPICE  
177 outperforms the current best method for rare and ultra-rare variants with various molecular effects.

178 In this study, we used the same set of features as CADD used for constructing their score  
179 but trained the model directly on pathogenicity. The features enabled CAPICE to make predictions  
180 for variants of various molecular effects. Its focus on pathogenicity helped CAPICE to overcome  
181 the challenges faced by CADD in predicting pathogenicity (35) in the clinic. As a result, CAPICE  
182 gives significantly better prediction for rare variants, and various types of variants, in particular,  
183 frameshift, splicing, and stop-gained variants. We also observed that most current predictors have  
184 problems classifying rare and ultra-rare variants, with the exception for REVEL, an ensemble  
185 method that targets rare variants. We thus adopted the same strategy as REVEL by including  
186 rare variants when training CAPICE, and thereby obtained a comparable performance to that of

187 REVEL for missense rare variants and significantly better results than all the other methods tested  
188 for ultra-rare variants.

189 We made full use of the large amount of data generated by other researchers. The  
190 evidence for a variant's clinical relevance reported in public databases such as ClinVar can be  
191 conflicting or outdated (36). The star system used in ClinVar review status (37) serves as a good  
192 quality check for estimating the trustworthiness of the reported pathogenicity, and this quality  
193 estimation is used by many researchers as a selection criteria for constructing or evaluating  
194 variant prioritization methods [(15), (38)]. However, this method of data selection can introduce  
195 biases and waste potentially important information. In particular, neutral variants can be enriched  
196 for common ones. These common variants can be easily filtered out in a diagnostic pipeline using  
197 a general cut-off or expected carrier prevalence for specific diseases (39). Using such a biased  
198 dataset could however lead to a biased model or an overly optimistic performance estimation.  
199 When training CAPICE, we did not exclude lower-quality data, and assigned it a lower sample  
200 weight during model training. This strategy overcome the data selection bias mentioned above  
201 and led to a model with equally good performance for rare and ultra-rare variants. When testing  
202 CAPICE, we only selected high-quality data for the pathogenic set. For the neutral set, we  
203 included rare and ultra-rare variants for all the types of variations found in general population  
204 studies (after filtering for known pathogenic variations and inheritance mode). This allowed us to  
205 avoid the bias discussed above.

206 Current variant prioritization methods, including ours, often neglect context information  
207 about a patient such as phenotype information, family history and the cell types associated with  
208 specific diseases. Moreover, the methods developed are often evaluated in a stand-alone manner,  
209 and their associations with other steps in a genome diagnostic pipeline are not often investigated.  
210 In this study, we have only shown preliminary evaluation results using solved patient data. In  
211 future studies, we hope to include context information to further improve CAPICE's predictive

212 power. We also believe that the model's performance needs to be discussed in a broader context  
213 that includes gene prioritization and mutational burden-testing.

214

## 215 **CONCLUSION**

216 We have developed CAPICE, an ensemble method for prioritizing pathogenic variants in clinical  
217 exomes for Mendelian disorders, including SNVs and InDels, that outperforms all other existing  
218 methods and that we dream will greatly benefit rare disease research and patients worldwide. By  
219 re-using the CADD features, but training a machine-learning model on variants' pathogenicity,  
220 CAPICE consistently outperforms other methods in our benchmark datasets for variants of  
221 various molecular effect and allele frequency. Additionally, we demonstrate that predictions made  
222 using CAPICE scores produce many fewer false positives than predictions made based on CADD  
223 scores. To enable its integration into automated and manual diagnostic pipelines, CAPICE is  
224 available as a free and open source software command-line tool from  
225 <https://github.com/molgenis/capice> and as a web-app at <https://molgenis43.gcc.rug.nl/>. Pre-  
226 computed scores are available as a download at <https://doi.org/10.5281/zenodo.3516248>.

227

## 228 **MATERIALS AND METHODS**

229 The flowchart of this study is in Supplementary Figure 1.

230

## 231 **DATA**

### 232 **Data collection and selection**

233 Training and benchmark data on neutral and pathogenic variants were derived from vcf files from  
234 the ClinVar database (17), dated 02 January 2019; from the VKGL data share consortium (40);  
235 from the GoNL data (41) and from data used in a previous study (29). From the ClinVar dataset,  
236 we collected variants reported by one or more submitters to have clear clinical significance,

237 including pathogenic and likely pathogenic variants and neutral and likely neutral variants. From  
238 the VKGL data consortium, we collected variants with clear classifications, either (Likely)  
239 Pathogenic or (Likely) Benign, with support from one or more laboratories. The neutral variants  
240 from previous research developing the GAVIN tool (29) were mainly collected from ExAC without  
241 posing a constraint on allele frequency. We also obtained a neutral benchmark dataset from a  
242 benchmark study by (24).

243 In our data selection step, we removed duplicate variants located in unique chromosomal  
244 positions and those with inconsistent pathogenicity classification across the different databases.  
245 To reduce potential variants in general population datasets from carriers, we excluded variants  
246 observed in dominant genes using inheritance modes of each gene retrieved from the Clinical  
247 Genome Database dated 28 February, 2019 (42).

248 In total, we collected 80k pathogenic variants and 450k putative neutral variants, and the  
249 entire dataset can be found in the Supplementary Material. After the initial cleaning step described  
250 above, we built a training dataset for model construction and a benchmark dataset that we left out  
251 of the training procedures so it could be used for performance evaluation later on.

252

### 253 **Construction of the benchmark and training sets**

254 To build a benchmark dataset for performance evaluation that was fully independent of model  
255 construction procedures, we first selected the high-confidence pathogenic variants from the  
256 ClinVar and VKGL database. High-confidence variants are those with a review status of “two or  
257 more submitters providing assertion criteria provided the same interpretation (criteria provided,  
258 multiple submitters, no conflicts)”, “review by expert panel” and “practice guideline” in ClinVar  
259 database, and those are reported by one of more laboratories without conflicting interpretation in  
260 VKGL database. From the pathogenic variants that passed these criteria, we then randomly  
261 selected 50% to add into the benchmark dataset, which resulted in 6,937 pathogenic variants. To

262 enable unbiased comparison of neutral and pathogenic variants with different molecular  
263 consequences, we created benchmark datasets with equal proportions of pathogenic and neutral  
264 variants for each type of molecular consequences, with the additional requirement that the  
265 pathogenic and neutral variants share similar distributions in allele frequency. An overview of the  
266 allele frequency distribution of the pathogenic and neutral variants for each type of molecular  
267 effects is in Supplementary Figure 2.

268 In total, our benchmark set contained 10,842 variants and our training set contained  
269 334,601 variants. The training set had 32,783 high confidence variants and 301,819 lower  
270 confidence variants. The high-confidence training variants were 12,646 pathogenic variants and  
271 20,137 neutral variants. The lower confidence variants were 28,035 pathogenic variants and  
272 273,784 neutral variants.

273 The two neutral benchmark datasets are those taken from a previous benchmark study  
274 and the GoNL dataset. The previous benchmark study by [(24)] selected neutral variants from the  
275 ExAC dataset and only included common variants with allele frequencies between 1% and 25%.  
276 For this dataset, we removed variants seen in the training set. In total, there were 60,699 neutral  
277 variants in our benchmark dataset. To build the neutral benchmark dataset from GoNL data, we  
278 selected all the variants that passed our quality assessment, then calculated their allele frequency  
279 within the GoNL population. We then selected those variants with an allele frequency  $<1\%$  and  
280 removed variants that had been included in the training set. In total, there were 14,426,914  
281 variants involved (Supplementary Table 2).

282

### 283 **Data annotation and preprocessing**

284 The collected variants in both the training and test datasets were annotated using CADD web  
285 service v1.4, which consists of 92 different features from VEP (version 90.5) (43) and epigenetic  
286 information from ENCODE (44) and the NIH RoadMap project (45). A detailed explanation of  
287 these features can be found in the (21) CADD paper. For each of the 11 categorical features, we

288 selected up to five top levels to avoid introducing excessive sparsity, which could be  
289 computationally expensive, and used one-hot encoding before feeding the data into the model  
290 training procedures (46). For the 81 numerical variables, we imputed each feature using the  
291 imputation value recommended by (21). The allele frequency in the population was annotated  
292 using the vcfTool (47) from GnomAD r2.0.1 (48). We assigned variants not found in the GnomAD  
293 database an allele frequency of 0.

294

## 295 **MODEL CONSTRUCTION**

### 296 **Model construction and training procedures**

297 We trained a gradient-boosting tree model using the XGBoost (version 0.72) Python package.  
298 The hyper-parameters, `n_estimators`, `max_depth` and `learning_rate` were selected by 5-fold  
299 cross-validation using the `RandomSearchCV` function provided by the `scikit-learn` (version 0.19.1)  
300 Python package. Within each training fold, we used an early stopping criteria of 15 iterations. We  
301 then used the model trained with the best set of hyper-parameters (0.1 for `learning_rate`, 15 for  
302 `max_depth` and 422 for `n_estimators`) for performance measurement. For fitting the model, we  
303 also used the sample weight assigned to each variant. The sample weight is a score ranging from  
304 0 to 1 that reflects the confidence level of the trustworthiness of the pathogenicity status of that  
305 variant. High-confidence variant, as described previously, are given a sample weight of 1, and the  
306 low-confidence variants were given a lower sample weight of 0.8. A variant with a high sample  
307 weight will thus contribute more to the loss function used in the training procedure (46). To test  
308 the assigned sample weights, we used the best set of parameters returned from the previous fine-  
309 tuning process and tried three different conditions in which we set the sample weights of the lower  
310 confidence variants to 0, 0.8 and 1. We then selected the model with the highest AUC value for  
311 the cross-validation dataset.

312

## 313 **Threshold-selection Strategies**

314 For comparing the false positive rate in the neutral benchmark dataset and comparing the  
315 classification results, we tested different threshold-selection strategies for both CAPICE and  
316 CADD. For CAPICE, we obtained the threshold from the training dataset that results in a recall  
317 value within 0.94-0.96. To calculate the threshold, we searched for all possible threshold value  
318 from 0 to 1 and selected the first threshold for which the resulting recall value fall between 0.94  
319 and 0.96. This method resulted in a general threshold of 0.02. For CADD, we tested two different  
320 threshold-selection methods. The first threshold was a default value of 20. The second method  
321 used GAVIN (29) to provide gene-specific thresholds. For other machine learning methods that  
322 returned a pathogenicity score ranging from 0 to 1, and no recommended threshold was given in  
323 the original paper, we selected a default value of 0.5. This includes the following methods: REVEL,  
324 ClinPred, SIFT and FATHMM-XF. For PROVEAN, we used a default score of -2.5 as the  
325 threshold.

326

## 327 **EVALUATION METRICS**

328 For model performance comparison, we used Receiver Operating Characteristic (ROC) curve,  
329 AUC value (49), and measurements in the confusion matrix together with the threshold-selection  
330 strategies mentioned above. For measuring model performance in the neutral benchmark dataset,  
331 we examined the false positive rate. The false positive rate is the number of true neutral variants  
332 but predicted as pathogenic divided by the number of true neutral variants. To evaluate the  
333 robustness of the model predictions, we performed bootstrap on the benchmark dataset for  
334 standard deviation measurement for 100 repetitions, with the same sample size of the benchmark  
335 dataset for each repetition (50).

336 For evaluating performance in solved patients, we used the previously diagnosed patients  
337 with clear record of the disease-causing variant from University Medical Center in Groningen. A

338 description of the solved patients' can be found in (34). For examining CAPICEs performance, we  
339 first eliminated all variants with an allele frequency above 10% and then predicted the  
340 pathogenicity for the remaining variants. Subsequently, we sorted the variants of each individual  
341 by their pathogenicity score assigned by the respective predictors, and used the ranking of the  
342 disease-causing variant found within that individual as the measurement.

343         The data and pathogenicity predictions are provided in Web resources.

344

### 345 **WEB RESOURCE**

346 CAPICE's Precomputed Scores: <https://doi.org/10.5281/zenodo.3516248>

347 CAPICE: <https://github.com/molgenis/capice> and web application <https://molgenis43.gcc.rug.nl>

348 CADD: <https://cadd.gs.washington.edu/score>

349 REVEL: <https://sites.google.com/site/revelgenomics/>

350 PON-P2: <http://structure.bmc.lu.se/PON-P2/>

351 ClinPred: <https://sites.google.com/site/clinpred/>

352 PROVEAN and SIFT: [http://provean.jcvi.org/genome\\_submit\\_2.php?species=human](http://provean.jcvi.org/genome_submit_2.php?species=human)

353 GAVIN: <https://molgenis.org/gavin>

354 FATHMM-XF: <http://fathmm.biocompute.org.uk/fathmm-xf/>

355

### 356 **DECLARATIONS**

357 All manuscripts must contain the following sections under the heading 'Declarations':

- 358         • Ethics approval and consent to participate

359         Not Applicable

- 360         • Consent for publication

361         Not Applicable

- 362         • Availability of data and materials



363 Training and testing data with label and predictions from CAPICE and tested predictors and  
364 the pre-computed scores for all possible SNVs and InDels is available online at Zenodo:  
365 <https://doi.org/10.5281/zenodo.3516248> and at Github: <https://github.com/molgenis/capice>.

366 • Competing interests

367 The authors declare that they have no competing interests

368 • Funding

369 This project has received funding from the Netherlands Organisation for Scientific Research  
370 NWO under VIDI grant number 917.164.455.

371 • Authors' contributions

372 Shuang Li, K. Joeri van der Velde, Morris A. Swertz designed the experiments, analyzed the  
373 data, and wrote the paper. K, Joeri van der Velde and Morris A. Swertz provided support in  
374 supervising Shuang Li in conducting the projects. Dick de Ridder, Aalt-Jan van Dijk, Dimitrios  
375 Soudis, and Leslie Zwerwer provided support for experimental design and model construction.  
376 Patrick Deelen provided support for experiment design and evaluation of the model in real  
377 patients' data. Dennis Hendriksen and Bart Charbon provided support for web application and  
378 web service API construction. Marielle van Gijn and Richard Sinke provided support for  
379 interpreting the results. Kristin Abbot, Birgit Sikkema, Cleo van Diemen, Mieke Kerstjens-  
380 Frederikse provided support in collecting the patients' diagnostic records and interpreting the  
381 results. All authors read and approved the final manuscript.

382 • Acknowledgements

383 Kate Mc Intyre contributed greatly in the development and refinement of texts. Harm-Jan  
384 Westra helped us in reviewing the manuscript. Tommy de Boer provided support with the web  
385 service API construction.

## 386 REFERENCES

387 1. Boudellioua I, Mahamad Razali RB, Kulmanov M, Hashish Y, Bajic VB, Goncalves-Serra  
388 E, et al. Semantic prioritization of novel causative genomic variants. PLoS Comput Biol

- 389 [Internet]. 2017 Apr [cited 2018 May 3];13(4):e1005500. Available from:  
390 <http://www.ncbi.nlm.nih.gov/pubmed/28414800>
- 391 2. Lionel AC, Costain G, Monfared N, Walker S, Reuter MS, Hosseini SM, et al. Improved  
392 diagnostic yield compared with targeted gene sequencing panels suggests a role for  
393 whole-genome sequencing as a first-tier genetic test. *Genet Med* [Internet]. 2018 Apr 3  
394 [cited 2018 May 9];20(4):435–43. Available from:  
395 <http://www.nature.com/doi/10.1038/gim.2017.119>
- 396 3. Clark MM, Hildreth A, Batalov S, Ding Y, Chowdhury S, Watkins K, et al. Diagnosis of  
397 genetic diseases in seriously ill children by rapid whole-genome sequencing and  
398 automated phenotyping and interpretation. *Sci Transl Med* [Internet]. 2019 Apr 24 [cited  
399 2019 Oct 2];11(489):eaat6177. Available from:  
400 <http://www.ncbi.nlm.nih.gov/pubmed/31019026>
- 401 4. Sawyer SL, Hartley T, Dymont DA, Beaulieu CL, Schwartzentruber J, Smith A, et al.  
402 Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time  
403 to address gaps in care. *Clin Genet* [Internet]. 2016 Mar [cited 2019 Oct 2];89(3):275–84.  
404 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26283276>
- 405 5. Trujillano D, Bertoli-Avella AM, Kumar Kandaswamy K, Weiss ME, Köster J, Marais A, et  
406 al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur J*  
407 *Hum Genet* [Internet]. 2017 Feb 16 [cited 2018 Nov 30];25(2):176–82. Available from:  
408 <http://www.nature.com/articles/ejhg2016146>
- 409 6. Meng L, Pammi M, Saronwala A, Magoulas P, Ghazi AR, Vetrini F, et al. Use of Exome  
410 Sequencing for Infants in Intensive Care Units. *JAMA Pediatr* [Internet]. 2017 Dec 4 [cited  
411 2019 Oct 2];171(12):e173438. Available from:  
412 <http://www.ncbi.nlm.nih.gov/pubmed/28973083>

- 413 7. Bardakjian TM, Helbig I, Quinn C, Elman LB, McCluskey LF, Scherer SS, et al. Genetic  
414 test utilization and diagnostic yield in adult patients with neurological disorders. [cited  
415 2018 Nov 30]; Available from: <https://doi.org/10.1007/s10048-018-0544-x>
- 416 8. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian  
417 disease. *Nat Rev Genet* [Internet]. 2017 Aug 14 [cited 2018 Jan 31];18(10):599–612.  
418 Available from: <http://www.nature.com/doi/10.1038/nrg.2017.52>
- 419 9. Thiffault I, Farrow E, Zellmer L, Berrios C, Miller N, Gibson M, et al. Clinical genome  
420 sequencing in an unbiased pediatric cohort. *Genet Med* [Internet]. 2019 Feb 16 [cited  
421 2019 Oct 2];21(2):303–10. Available from: [http://www.nature.com/articles/s41436-018-](http://www.nature.com/articles/s41436-018-0075-8)  
422 [0075-8](http://www.nature.com/articles/s41436-018-0075-8)
- 423 10. Berberich AJ, Ho R, Hegele RA. Whole genome sequencing in the clinic: empowerment  
424 or too much information? *CMAJ* [Internet]. 2018 [cited 2019 Oct 2];190(5):E124–5.  
425 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29431109>
- 426 11. Shi F, Yao Y, Bin Y, Zheng C-H, Xia J. Computational identification of deleterious  
427 synonymous variants in human genomes using a feature-based approach. *BMC Med*  
428 *Genomics* [Internet]. 2019 Jan 31 [cited 2019 Oct 2];12(S1):12. Available from:  
429 <https://bmcmmedgenomics.biomedcentral.com/articles/10.1186/s12920-018-0455-6>
- 430 12. Jagadeesh KA, Paggi JM, Ye JS, Stenson PD, Cooper DN, Bernstein JA, et al. S-CAP  
431 extends pathogenicity prediction to genetic variants that affect RNA splicing. *Nat Genet*  
432 [Internet]. 2019 Apr 25 [cited 2019 Oct 2];51(4):755–63. Available from:  
433 <http://www.nature.com/articles/s41588-019-0348-4>
- 434 13. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF:  
435 accurate prediction of pathogenic point mutations via extended features. Hancock J,  
436 editor. *Bioinformatics* [Internet]. 2018 Feb 1 [cited 2019 Oct 2];34(3):511–3. Available  
437 from: <http://www.ncbi.nlm.nih.gov/pubmed/28968714>

- 438 14. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function.  
439 Nucleic Acids Res [Internet]. 2003 Jul 1 [cited 2019 Oct 2];31(13):3812–4. Available from:  
440 <http://www.ncbi.nlm.nih.gov/pubmed/12824425>
- 441 15. Alirezaie N, Kernohan KD, Hartley T, Majewski J, Hocking TD. ClinPred: Prediction Tool  
442 to Identify Disease-Relevant Nonsynonymous Single-Nucleotide Variants. Am J Hum  
443 Genet [Internet]. 2018 Oct 4 [cited 2019 Oct 2];103(4):474–83. Available from:  
444 <http://www.ncbi.nlm.nih.gov/pubmed/30220433>
- 445 16. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL:  
446 An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J  
447 Hum Genet [Internet]. 2016 Oct 6 [cited 2019 Oct 2];99(4):877–85. Available from:  
448 <http://www.ncbi.nlm.nih.gov/pubmed/27666373>
- 449 17. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar:  
450 public archive of relationships among sequence variation and human phenotype. Nucleic  
451 Acids Res [Internet]. 2014 Jan [cited 2019 Oct 2];42(Database issue):D980-5. Available  
452 from: <http://www.ncbi.nlm.nih.gov/pubmed/24234437>
- 453 18. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al.  
454 Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes.  
455 Genome Res [Internet]. 2005 Aug 1 [cited 2019 Oct 2];15(8):1034–50. Available from:  
456 <http://www.ncbi.nlm.nih.gov/pubmed/16024819>
- 457 19. Davydov E V., Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High  
458 Fraction of the Human Genome to be under Selective Constraint Using GERP++.  
459 Wasserman WW, editor. PLoS Comput Biol [Internet]. 2010 Dec 2 [cited 2019 Oct  
460 2];6(12):e1001025. Available from: <https://dx.plos.org/10.1371/journal.pcbi.1001025>
- 461 20. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the  
462 pathogenicity of genetic variants. Bioinformatics [Internet]. 2015 Mar 1 [cited 2019 Oct

- 463 2];31(5):761–3. Available from: [https://academic.oup.com/bioinformatics/article-](https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu703)  
464 [lookup/doi/10.1093/bioinformatics/btu703](https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu703)
- 465 21. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework  
466 for estimating the relative pathogenicity of human genetic variants. *Nat Genet* [Internet].  
467 2014 Mar 2 [cited 2019 Oct 2];46(3):310–5. Available from:  
468 <http://www.nature.com/articles/ng.2892>
- 469 22. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the  
470 deleteriousness of variants throughout the human genome. *Nucleic Acids Res* [Internet].  
471 2019 Jan 8 [cited 2019 Oct 2];47(D1):D886–94. Available from:  
472 <https://academic.oup.com/nar/article/47/D1/D886/5146191>
- 473 23. Fu W, O’Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, et al. Analysis of 6,515  
474 exomes reveals the recent origin of most human protein-coding variants. *Nature*  
475 [Internet]. 2013 Jan 28 [cited 2019 Oct 2];493(7431):216–20. Available from:  
476 <http://www.nature.com/articles/nature11690>
- 477 24. Niroula A, Vihinen M. How good are pathogenicity predictors in detecting benign  
478 variants? Panchenko ARR, editor. *PLOS Comput Biol* [Internet]. 2019 Feb 11 [cited 2019  
479 Oct 2];15(2):e1006481. Available from: <http://dx.plos.org/10.1371/journal.pcbi.1006481>
- 480 25. Ghosh R, Oak N, Plon SE. Evaluation of in silico algorithms for use with ACMG/AMP  
481 clinical variant interpretation guidelines. *Genome Biol* [Internet]. 2017 Dec 28 [cited 2018  
482 Jan 15];18(1):225. Available from:  
483 <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-017-1353-5>
- 484 26. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and  
485 integration of deleteriousness prediction methods for nonsynonymous SNVs in whole  
486 exome sequencing studies. *Hum Mol Genet* [Internet]. 2015 Apr 15 [cited 2018 May  
487 7];24(8):2125–37. Available from: [https://academic.oup.com/hmg/article-](https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddu733)  
488 [lookup/doi/10.1093/hmg/ddu733](https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddu733)

- 489 27. Schaafsma GCP, Vihinen M. VariSNP, A Benchmark Database for Variations From  
490 dbSNP. *Hum Mutat* [Internet]. 2015 Feb [cited 2019 Oct 2];36(2):161–6. Available from:  
491 <http://doi.wiley.com/10.1002/humu.22727>
- 492 28. Sarkar A, Yang Y, Vihinen M. Variation Benchmark Datasets: Update, Criteria, Quality  
493 and Applications. *bioRxiv* [Internet]. 2019 May 10 [cited 2019 Oct 2];634766. Available  
494 from: <https://www.biorxiv.org/content/10.1101/634766v1>
- 495 29. van der Velde KJ, de Boer EN, van Diemen CC, Sikkema-Raddatz B, Abbott KM,  
496 Knopperts A, et al. GAVIN: Gene-Aware Variant INterpretation for medical sequencing.  
497 *Genome Biol* [Internet]. 2017 Dec 16 [cited 2019 Oct 2];18(1):6. Available from:  
498 <http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1141-7>
- 499 30. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the  
500 deleteriousness of variants throughout the human genome. *Nucleic Acids Res* [Internet].  
501 2019 Jan 8 [cited 2019 Oct 2];47(D1):D886–94. Available from:  
502 <http://www.ncbi.nlm.nih.gov/pubmed/30371827>
- 503 31. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants  
504 from functional and population genomic data. *Nat Genet* [Internet]. 2017 Mar 13 [cited  
505 2018 Jan 15];49(4):618–24. Available from:  
506 <http://www.nature.com/doi/10.1038/ng.3810>
- 507 32. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional  
508 genomic annotations for coding and noncoding variants. *Nat Genet* [Internet]. 2016 Feb  
509 [cited 2019 Oct 23];48(2):214–20. Available from:  
510 <http://www.ncbi.nlm.nih.gov/pubmed/26727659>
- 511 33. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-  
512 based sequence model. *Nat Methods*. 2015 Sep 29;12(10):931–4.
- 513 34. Deelen P, van Dam S, Herkert JC, Karjalainen JM, Brugge H, Abbott KM, et al. Improving  
514 the diagnostic yield of exome- sequencing by predicting gene–phenotype associations

- 515 using large-scale gene expression analysis. *Nat Commun* [Internet]. 2019 Dec 28 [cited  
516 2019 Oct 2];10(1):2837. Available from: [http://www.nature.com/articles/s41467-019-](http://www.nature.com/articles/s41467-019-10649-4)  
517 10649-4
- 518 35. Mather CA, Mooney SD, Salipante SJ, Scroggins S, Wu D, Pritchard CC, et al. CADD  
519 score has limited clinical validity for the identification of pathogenic variants in noncoding  
520 regions in a hereditary cancer panel. *Genet Med* [Internet]. 2016 Dec 5 [cited 2019 Oct  
521 2];18(12):1269–75. Available from: <http://www.nature.com/articles/gim201644>
- 522 36. Shah N, Hou Y-CC, Yu H-C, Sainger R, Caskey CT, Venter JC, et al. Identification of  
523 Misclassified ClinVar Variants via Disease Population Prevalence. *Am J Hum Genet*  
524 [Internet]. 2018 Apr [cited 2019 Oct 2];102(4):609–19. Available from:  
525 <https://linkinghub.elsevier.com/retrieve/pii/S0002929718300879>
- 526 37. Review status in ClinVar [Internet]. [cited 2019 Oct 2]. Available from:  
527 [https://www.ncbi.nlm.nih.gov/clinvar/docs/review\\_status/](https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/)
- 528 38. Bao R, Huang L, Andrade J, Tan W, Kibbe WA, Jiang H, et al. Review of current  
529 methods, applications, and data management for the bioinformatics analysis of whole  
530 exome sequencing. *Cancer Inform* [Internet]. 2014 [cited 2018 Jan 19];13(Suppl 2):67–  
531 82. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25288881>
- 532 39. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and  
533 guidelines for the interpretation of sequence variants: a joint consensus recommendation  
534 of the American College of Medical Genetics and Genomics and the Association for  
535 Molecular Pathology. 2015 [cited 2018 Jan 15]; Available from:  
536 [https://www.acmg.net/docs/Standards\\_Guidelines\\_for\\_the\\_Interpretation\\_of\\_Sequence\\_](https://www.acmg.net/docs/Standards_Guidelines_for_the_Interpretation_of_Sequence_Variants.pdf)  
537 [Variants.pdf](https://www.acmg.net/docs/Standards_Guidelines_for_the_Interpretation_of_Sequence_Variants.pdf)
- 538 40. Fokkema IFAC, Velde KJ, Slofstra MK, Ruivenkamp CAL, Vogel MJ, Pfundt R, et al.  
539 Dutch genome diagnostic laboratories accelerated and improved variant interpretation  
540 and increased accuracy by sharing data. *Hum Mutat* [Internet]. 2019 Sep 3 [cited 2019

- 541 Oct 15];[humu.23896](https://doi.org/10.1002/humu.23896). Available from:  
542 <https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.23896>
- 543 41. Boomsma DI, Wijmenga C, Slagboom EP, Swertz MA, Karssen LC, Abdellaoui A, et al.  
544 The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* [Internet].  
545 2014 Feb [cited 2019 Oct 15];22(2):221–7. Available from:  
546 <http://www.nature.com/articles/ejhg2013118>
- 547 42. Solomon BD, Nguyen A-D, Bear KA, Wolfsberg TG. Clinical Genomic Database. *Proc*  
548 *Natl Acad Sci* [Internet]. 2013 Jun 11 [cited 2019 Oct 15];110(24):9851–5. Available from:  
549 <http://www.pnas.org/cgi/doi/10.1073/pnas.1302575110>
- 550 43. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl  
551 Variant Effect Predictor. *Genome Biol* [Internet]. 2016 Dec 6 [cited 2019 Oct 2];17(1):122.  
552 Available from: [http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-](http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4)  
553 [0974-4](http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4)
- 554 44. ENCODE Project Consortium TEP. An integrated encyclopedia of DNA elements in the  
555 human genome. *Nature* [Internet]. 2012 Sep 6 [cited 2019 Oct 2];489(7414):57–74.  
556 Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22955616>
- 557 45. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A,  
558 et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol* [Internet].  
559 2010 Oct [cited 2019 Oct 2];28(10):1045–8. Available from:  
560 <http://www.ncbi.nlm.nih.gov/pubmed/20944595>
- 561 46. Chen T, Guestrin C. XGBoost. In: *Proceedings of the 22nd ACM SIGKDD International*  
562 *Conference on Knowledge Discovery and Data Mining - KDD '16* [Internet]. New York,  
563 New York, USA: ACM Press; 2016 [cited 2019 Oct 2]. p. 785–94. Available from:  
564 <http://dl.acm.org/citation.cfm?doid=2939672.2939785>



- 565 47. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant  
566 call format and VCFtools. *Bioinformatics* [Internet]. 2011 Aug 1 [cited 2019 Oct  
567 2];27(15):2156–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21653522>
- 568 48. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. Variation  
569 across 141,456 human exomes and genomes reveals the spectrum of loss-of-function  
570 intolerance across human protein-coding genes. *bioRxiv* [Internet]. 2019 [cited 2019 Oct  
571 24];531210. Available from: <https://www.biorxiv.org/content/10.1101/531210v2>
- 572 49. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating  
573 characteristic (ROC) curve. *Radiology* [Internet]. 1982 Apr [cited 2019 Oct 2];143(1):29–  
574 36. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/7063747>
- 575 50. Bishop CM. *Pattern Recognition And Machine Learning* - Springer 2006. 2006.  
576

## 577 SUPPLEMENTARY MATERIALS

### 578 Additional Figures and Tables

Supplementary Figure 1 Flowchart of this study

Supplementary Figure 2 For benchmarking, we created a balanced test dataset that is equally distributed in terms of a) the number of pathogenic and putatively neutral variants and b) allele frequency distribution for different molecular consequences.

Supplementary Figure 3 ROC curves and AUC values from CAPICE and CADD for variants with different molecular functions.

Supplementary Figure 4 AUC and ROC curves for the a) full dataset and b) missense subset for rare and ultra-rare variants defined as variants with allele frequency between 0.01% and 0.1% and variants with allele frequency <0.01%

Supplementary Table 1 Description of all methods tested in the study

Method name	Application	Link for web resources
CADD	Estimating the relative pathogenicity of SNVs and InDels	<a href="https://cadd.gs.washington.edu/">https://cadd.gs.washington.edu/</a>
REVEL	Predicting the pathogenicity of rare missense variants	<a href="https://sites.google.com/site/revelgenomics/">https://sites.google.com/site/revelgenomics/</a>
ClinPred	Predicting the pathogenicity of missense variants	<a href="https://sites.google.com/site/clinpred/">https://sites.google.com/site/clinpred/</a>
PON-P2	Predicting the pathogenicity of missense variants	<a href="http://structure.bmc.lu.se/PON-P2/">http://structure.bmc.lu.se/PON-P2/</a>
SIFT	Predicting missense variants effects on protein function	<a href="http://provean.jcvi.org/genome_submit_2.php?species=human">http://provean.jcvi.org/genome_submit_2.php?species=human</a>
PROVEAN	Predicting missense variants and InDels' effects on protein function	<a href="http://provean.jcvi.org/genome_submit_2.php?species=human">http://provean.jcvi.org/genome_submit_2.php?species=human</a>

FATHMM-XF	Predicting pathogenicity of point mutations	<a href="http://fathmm.biocompute.org.uk/fathmm-xf/">http://fathmm.biocompute.org.uk/fathmm-xf/</a>
-----------	---	---

579

Supplementary Table 2 CAPICE and CADD false positive rates in the neutral benchmark dataset

Molecular Consequence	CADD (20)	CAPICE (General Threshold)	Number of Variants
NON_SYNONYMOUS	0.37	0.07	50181
DOWNSTREAM	0.36	0.11	12121
REGULATORY	0.35	0.12	11317
UPSTREAM	0.35	0.1	10959
INTRONIC	0.25	0.06	10910
NONCODING_CHANGE	0.34	0.06	1202
3PRIME_UTR	0.23	0.03	902
SYNONYMOUS	0.07	0.08	456
5PRIME_UTR	0.19	0.03	295
SPLICE_SITE	0.15	0.08	137
CANONICAL_SPLICE	0.6	0.4	10
STOP_GAINED	1	0.22	9

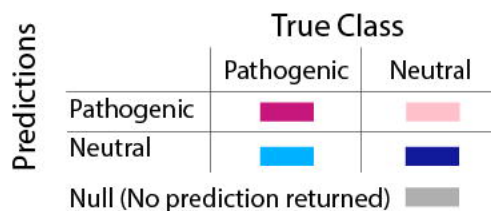
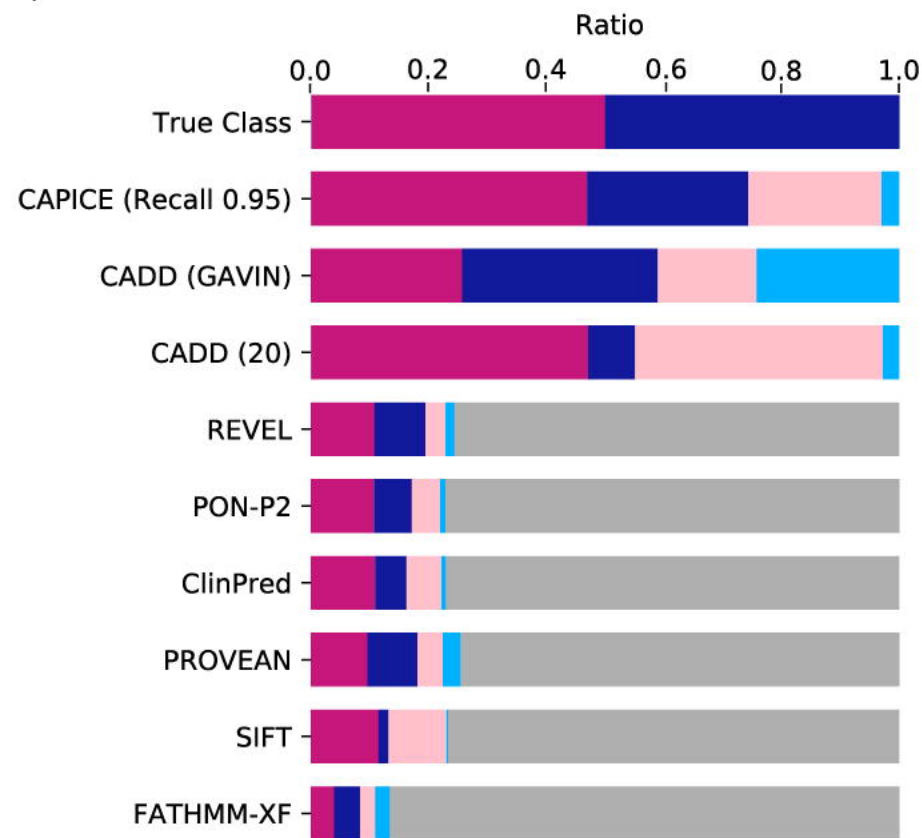
580

Supplementary Table 3 CAPICE and CADD false positive rates in the GoNL dataset

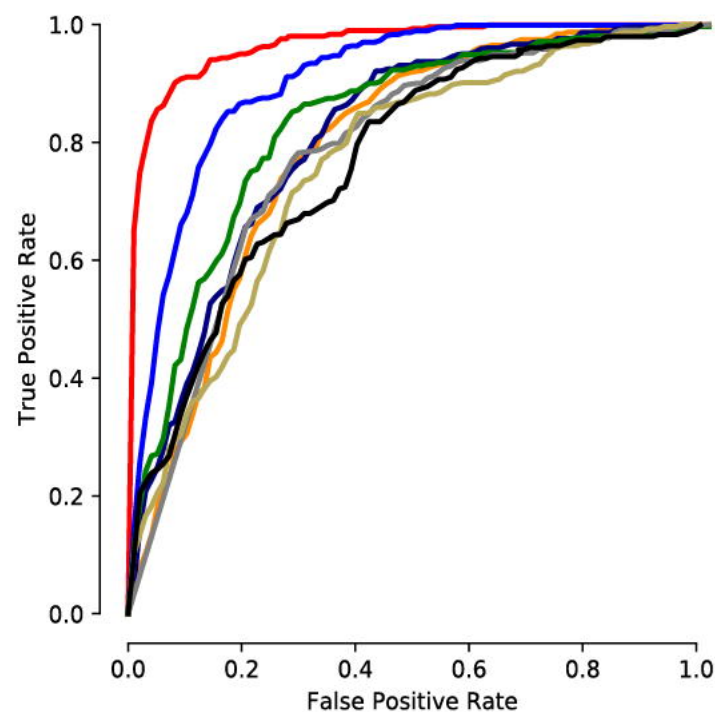
	CADD (20)	CAPICE	Number of Variants
INTRONIC	0.00	0.00	6483483
INTERGENIC	0.00	0.00	4399474

DOWNSTREAM	0.01	0.00	1205398
UPSTREAM	0.01	0.00	1174342
REGULATORY	0.01	0.01	808360
NONCODING_CHANGE	0.01	0.00	115099
3PRIME_UTR	0.02	0.00	108194
NON_SYNONYMOUS	0.57	0.19	67013
SYNONYMOUS	0.02	0.08	36177
5PRIME_UTR	0.03	0.01	14526
SPLICE_SITE	0.05	0.08	10352
CANONICAL_SPLICE	0.58	0.54	1545
STOP_GAINED	1.00	0.74	1647
FRAME_SHIFT	0.88	0.9	821
INFRAME	0.37	0.63	415
STOP_LOST	0.06	0.37	68

a) Variants of various molecular effects

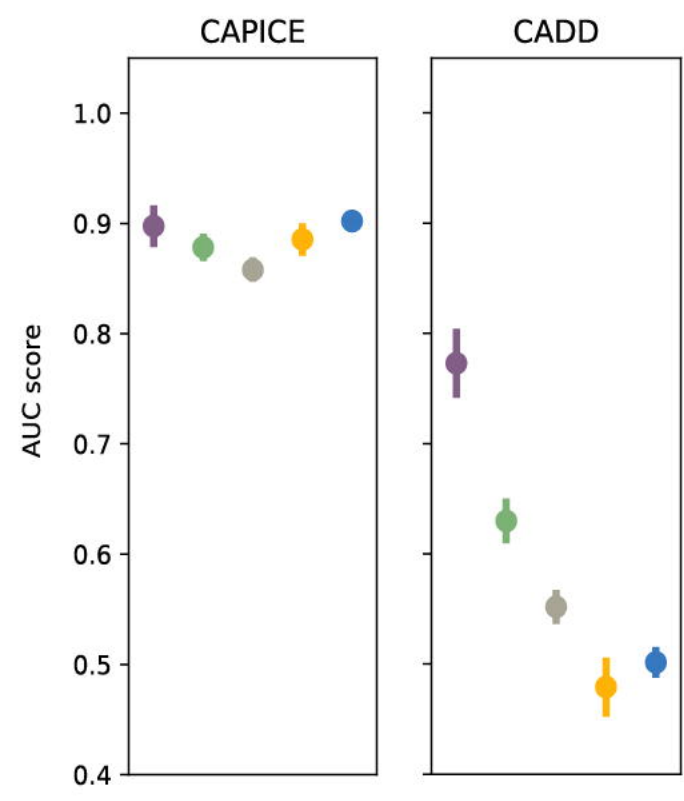


b) Non-synonymous variants



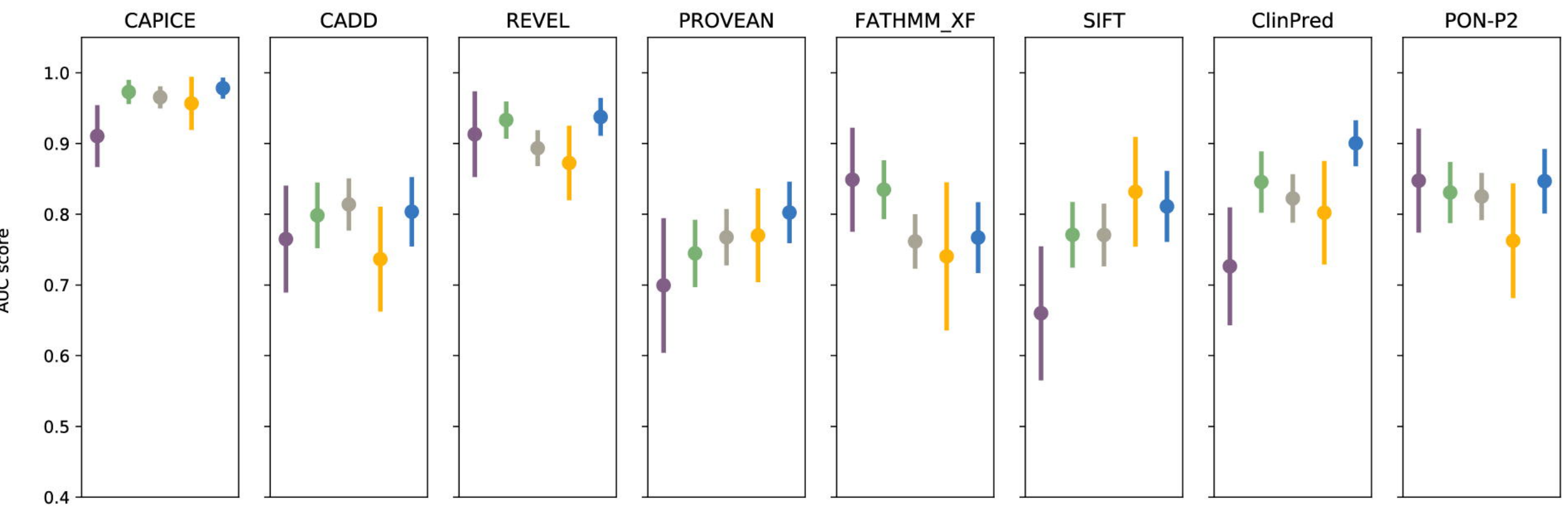
Method	AUC
CAPICE	0.97
CADD	0.79
REVEL	0.90
PON-P2	0.83
ClinPred	0.81
SIFT	0.78
PROVEAN	0.76
FATHMM-XF	0.77

a) Variants of various molecular effects

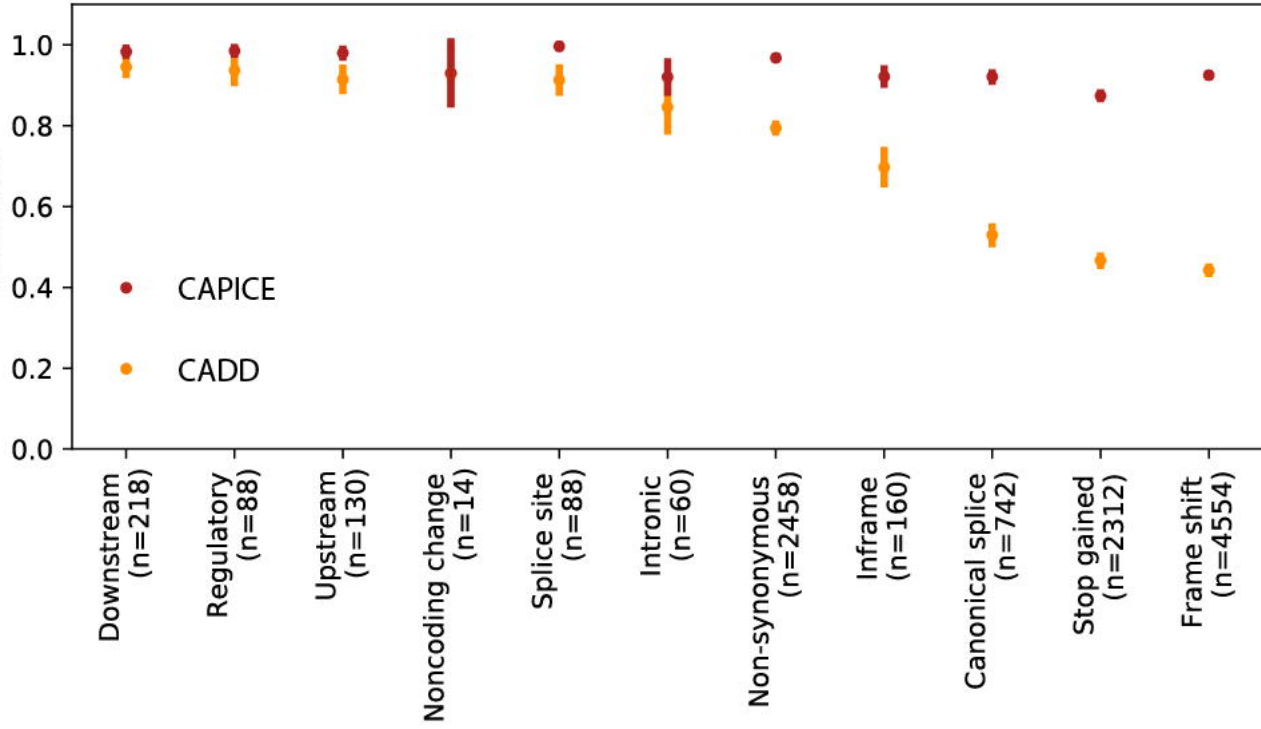


	Allele Frequency	Number of Variants in the full dataset	Number of Variants in non-synonymous subset
●	< 0.0001%	5089	195
●	0.001% ~ 0.0001%	1546	86
●	0.001% ~ 0.01%	2764	184
●	0.01% ~ 0.1%	1041	111
●	0.1% ~ 1%	374	39

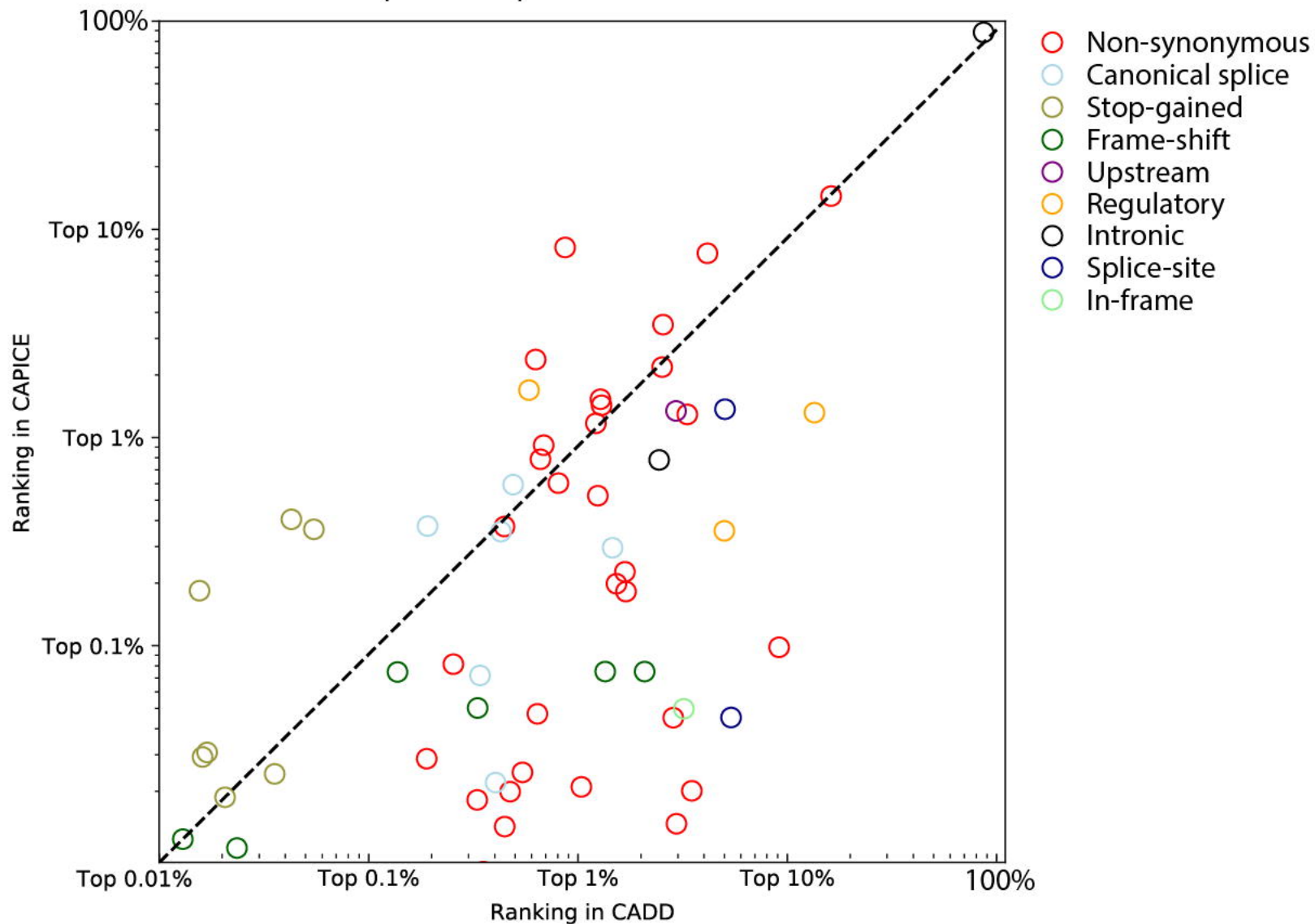
b) Non-synonymous variants



AUC score



Comparison of performance in real cases








VKGL + ClinVar + ExAC + dataset from another study

Train dataset

Test dataset\*

5-fold Cross-Validation for hyperparameter finetuning














 Test subset in the fold  
 Train subset in the fold  
 Subset of the entire training data for early-stopping

a) Split by Allele Frequency

$\leq 1e-6$   
 $1e-6 \sim 1e-5$   
 $1e-5 \sim 1e-4$   
 $1e-4 \sim 1e-3$   
 $1e-3 \sim 1e-2$   
 $> 0.01$

b) Split by Molecule Consequences

 Non-Synonymous  
 Canonical splicing  
 Stop-gained  
 Frame-shift  
 Upstream  
 Regulatory  
 Intronic  
 Splice site  
 In-frame  
 Downstream  
 Noncoding changes

Methods Compared

CAPICE  
CADD  
REVEL  
ClinPred  
PON-P2  
SIFT  
PROVEAN  
FATHMM-XF

CAPICE  
CADD

Other datasets for model performance comparison

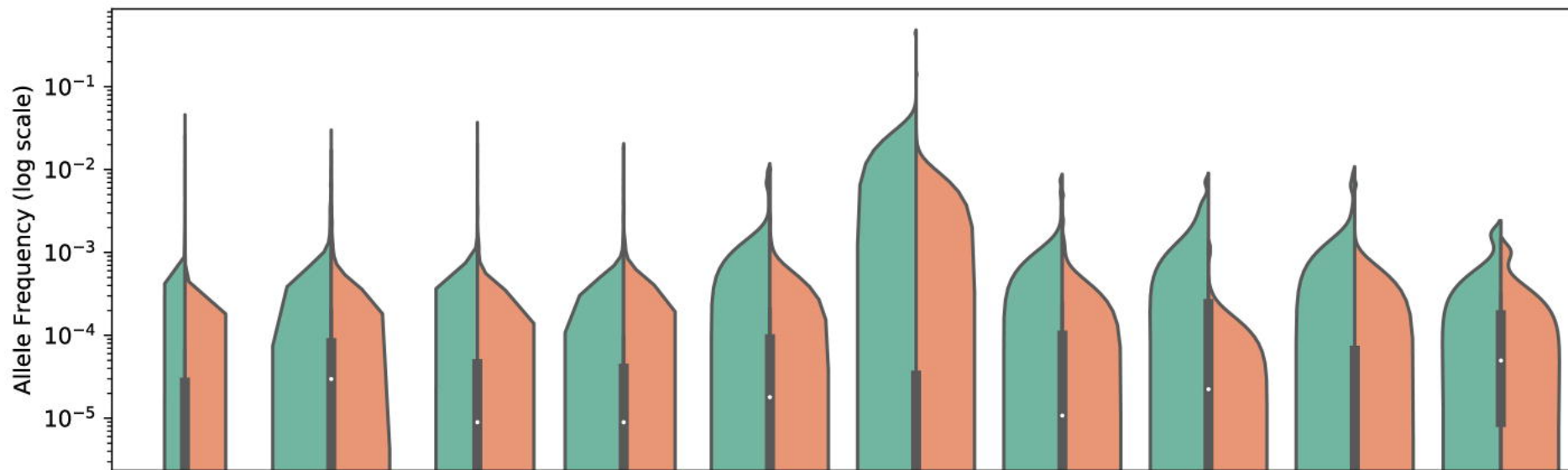
54 solved patients

GoNL as neutral set

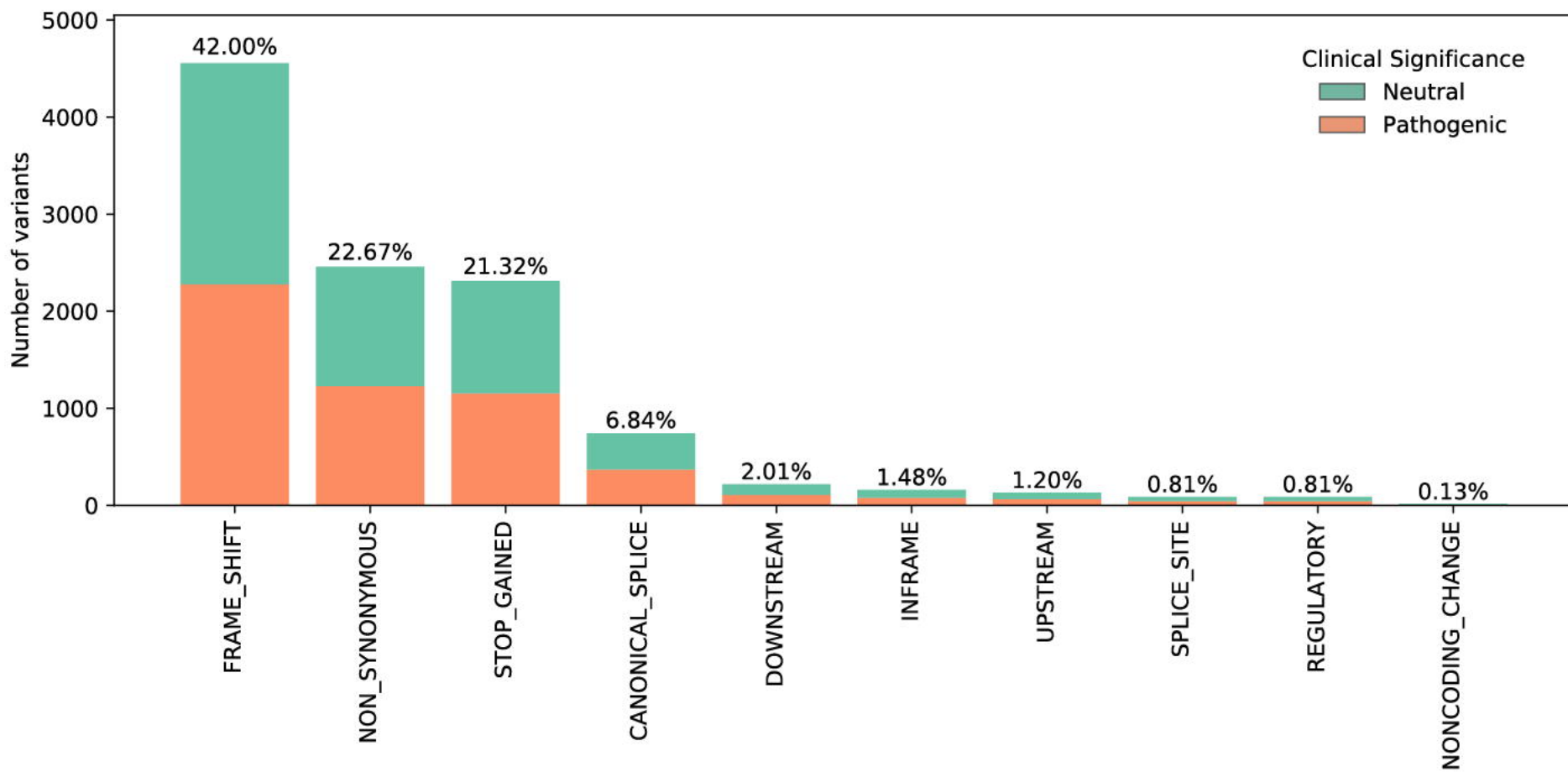
A neutral benchmark set in another study

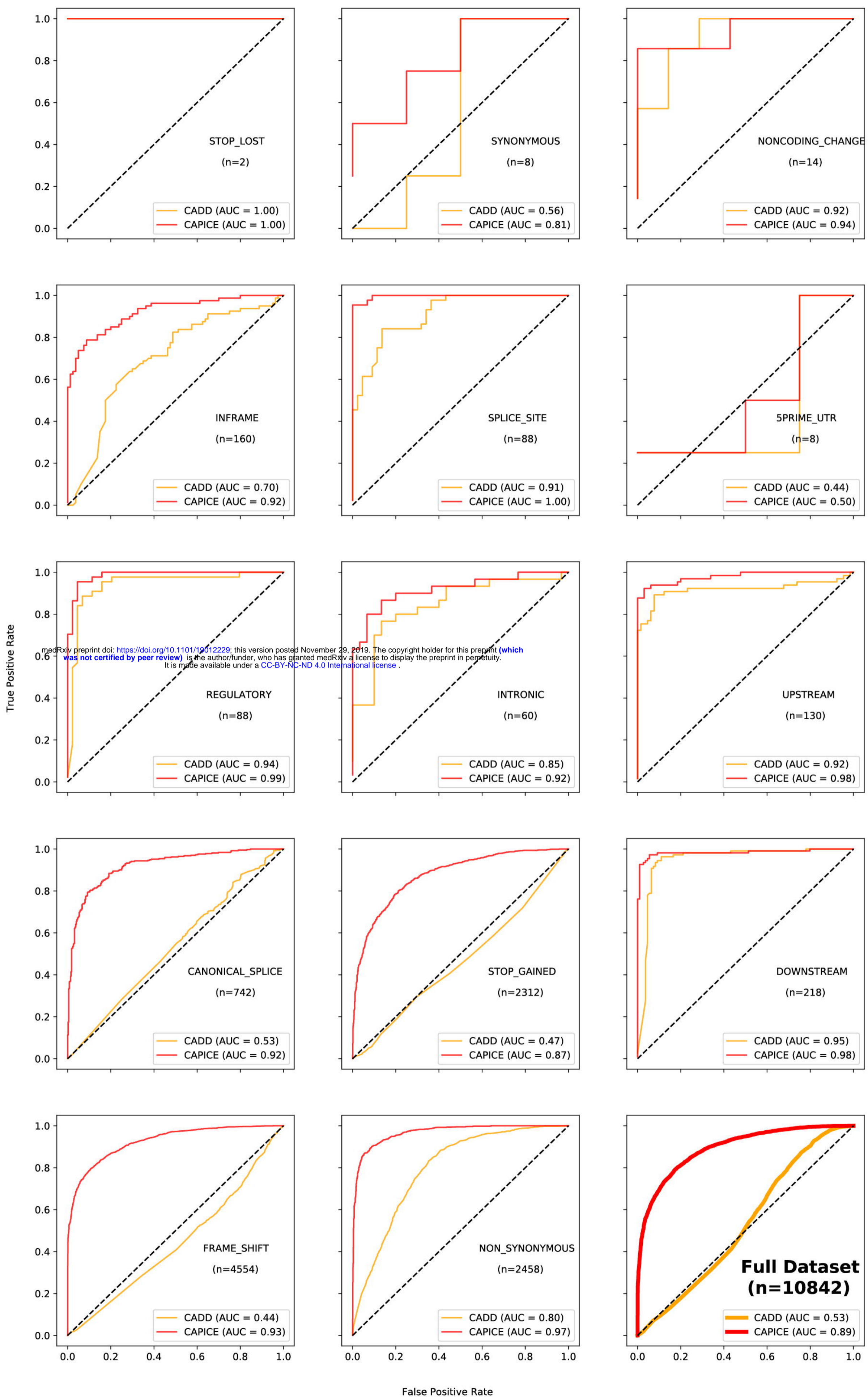
\* See Material and Methods for how the train and test datasets were split

a)

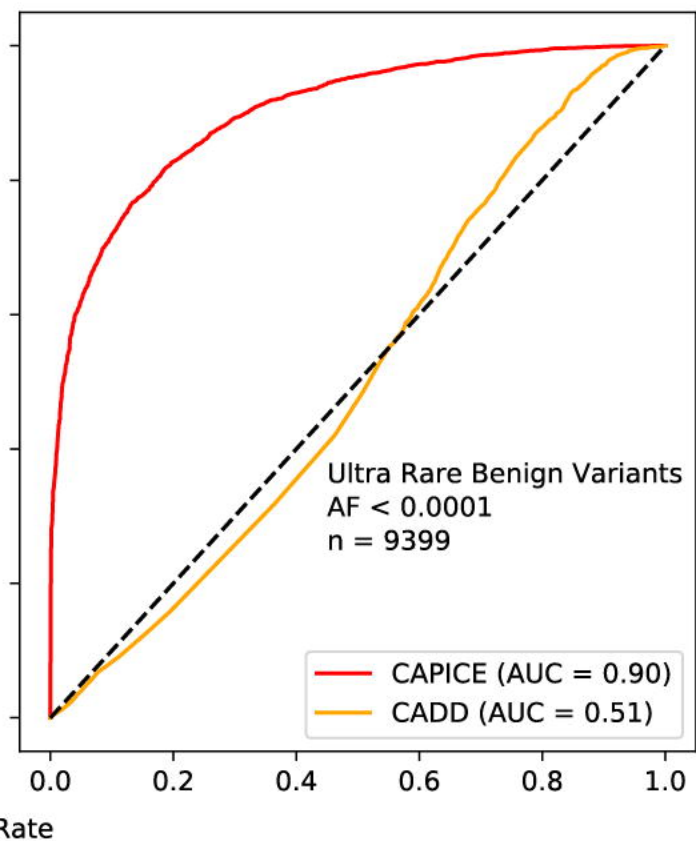
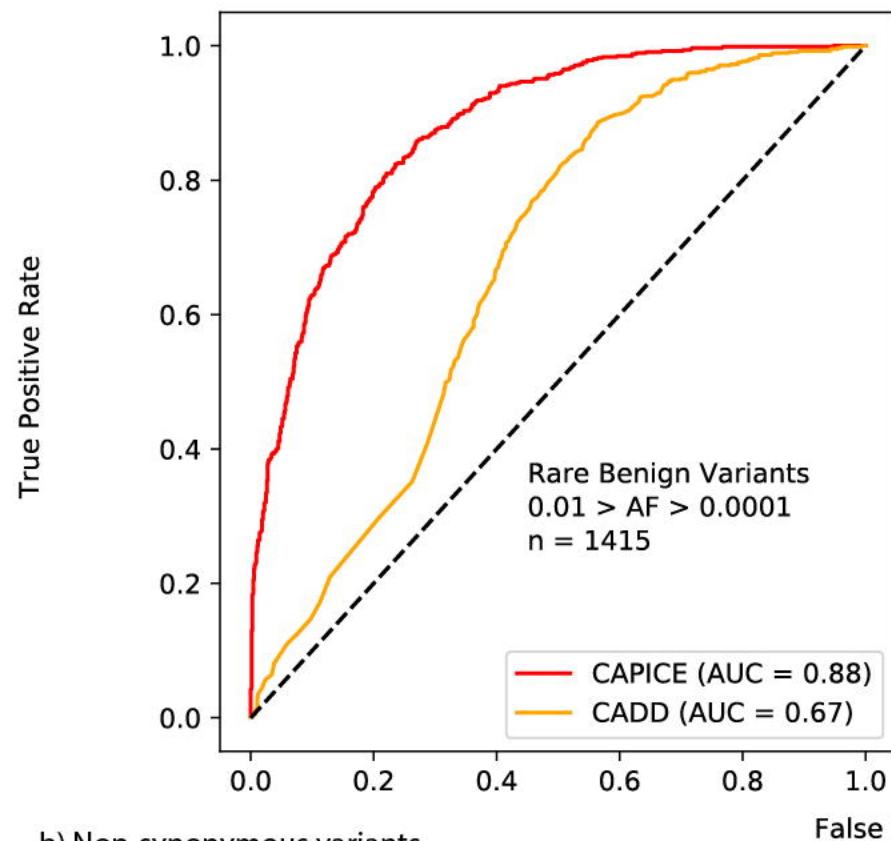


b)





a) Variants of all various molecular effects



b) Non-synonymous variants

